AD-A060 682 SYRACUSE UNIV N Y DEPT OF INDUSTRIAL ENGINEERING AND--ETC F/G 12/1
OPTIMAL CONTROL OF RANDOM WALKS, BIRTH AND DEATH PROCESSES AND --ETC(U)
1978 R F SERFOZO AFOSR-74-2627

UNCLASSIFIED AFOSR-TR-78-1345 NL

1 OF 1
AD
A060682

END
DATE
FILMED
1-79
DDC

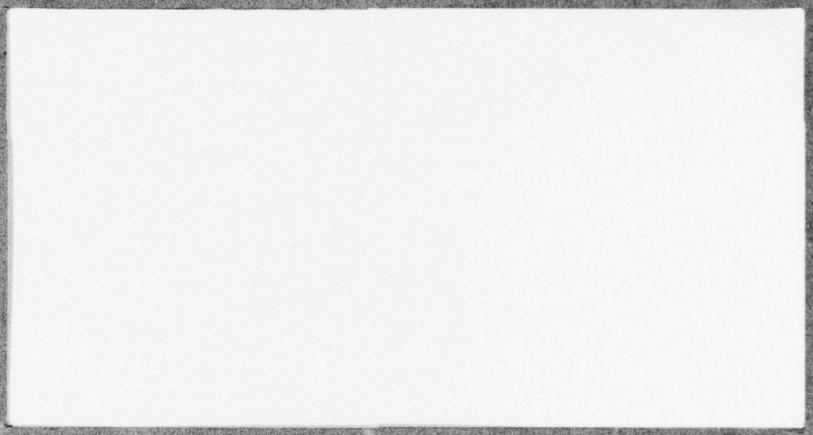
AFOSR-TR-78-1345

AD A060682

DDC FILE COPY

(2) LEVEL II

DEPARTMENT OF INDUSTRIAL ENGINEERING AND OPERATIONS RESEARCH



DDC
NOV 2 1978
B

COLLEGE OF ENGINEERING
SYRACUSE UNIVERSITY
SYRACUSE, NEW YORK 13210

78 09 13 109

FICE OF SCIENTIFIC RESEARCH (AFSC)
ARKSMITTAL TO DDC
al report has been reviewed and is
public release IAW AFR 190-12 (7b).
is unlimited.

nformation Officer

(2) LEVEL II

AD A060682

DDC FILE COPY

OPTIMAL CONTROL OF RANDOM WALKS, BIRTH AND DEATH PROCESSES AND QUEUES

by

RICHARD SERFOZO

DDC
RECEIVED
NOV 2 1978
B

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

78 09 13 109

ABSTRACT

We first consider a random walk on the nonnegative integers whose steps are controlled as follows. Upon arriving at a location i, a pair of probabilities (p,q) is selected from a given set, a reward r(i,p,q) is received, and the next step takes the walk to locations i+1, i-1 or i, with respective probabilities p, q and 1-p-q (q=0 when i=0). This is repeated indefinitely. A rule for successively selecting the probabilities (p,q) is a control policy. We identify conditions on the rewards and probabilities under which there exist monotone optimal policies, for discounted and average rewards, and we show how to compute some of these policies. For example, in certain settings it is optimal to increase the tendency of backward movement of the walk as its location increases. We also present similar results for controlling the parameters of a birth and death process and an M|M|1 queue.

Key Words: Random walk, birth and death process, M|M|1 queue, Markov decision process, dynamic programming.

RE: Classified reference, distribution unlimited-
No change per Ms. Blose, AFOSR

ACCESSION for
NTIS White Section
DDC Buff Section
UNANNOUNCED
JUSTIFICATION
BY
DISTRIBUTION/AVAILABILITY CODES
Dist. AVAIL. and/or SPECIAL
A

OPTIMAL CONTROL OF RANDOM WALKS, BIRTH AND DEATH PROCESSES AND QUEUES

by

Richard Serfozo
Syracuse University

1. Introduction

This study was motivated by the following queueing control problem. Consider a single server queueing system with exponential interarrival and service times (an M|M|1 queue) whose rates are controlled as follows. At each customer arrival or service completion, the number of customers in the system is observed. Based on this number, an arrival and service rate pair (λ,μ) is selected from a given set. Costs are incurred for using the rates (λ,μ), and for holding customers in the system. The problem is to find a policy for selecting the rates so as to minimize the (discounted or average) cost of running the system. It seems reasonable, that the optimal arrival and service rates should be decreasing and increasing functions, respectively, of the number of waiting customers. That is, as the queue increases, there should be faster service and fewer arrivals in order to reduce the queue. The question is what types of costs lead to such monotone optimal policies? We address this problem herein for a more general birth and death process and for a simple random walk.

We begin by considering a random walk on the nonnegative integers whose steps are of size $+1$, -1 or 0 with respective probabilities p, q, and $1-p-q$, where the (p,q) is

---

This research was partially sponsored by the Air Force Office of Scientific Research Grant #AFOSR-74-2627, and by the National Science Foundation Grant #ENG75-13653.

selected at each step depending on the location. In Sections 3-5, we present very general conditions under which it is optimal to increase (or decrease) the tendency of backward movement of the walk as its location increases. We do this for discounted and average rewards over infinite or finite time horizons. In Section 5 we discuss the case where the (p,q)'s may depend on the location. Our analysis is based on three results that apply to more general Markov decision processes. They are (1) a criterion for establishing the existence of a monotone optimal policy (Proposition 2.1, which is generalized in Serfozo (1977) and Topkis (1978)), (2) a sufficient condition for the upper envelope of a family of concave functions, defined on the integers, to be concave (Proposition 2.2), and (3) a result for obtaining a monotone average optimal policy from a set of monotone discounted optimal policies (Proposition 4.1, an extension of Theorem 1 in Derman (1962)).

In Section 6 we consider a birth and death process whose birth and death rates are controlled whenever the process takes a jump. We show that the optimal policies for this continuous time process are the same as those for a discrete time random walk as described above. This follows by an equivalence between continuous and discrete time Markov decision processes (see Lippman (1975) and Serfozo (1979)), which is an extension of that used by Howard and Veinott. We then apply our results for random walks to describe optimal policies for controlling the birth and death process. A special case is the above M|M|1 queue with controlled arrival and

service rates. Similar queueing control models are (1) M|M|1 queue with controlled service rate [1]-[3], [10] and [13], (2) M|M|1 queue with controlled arrival rate [10]-[12], and (3) M|G|1 queue with service switched on or off [5], [8] and [17]. Many other studies of controlled queues are referenced in [4], [18] and [19].

2. Preliminaries

The following are some basics of controlled Markov chains (discrete time Markov decision processes), that we shall use in our analysis.

Consider a controlled Markov chain that moves in the space of nonnegative integers as follows. Upon arriving at a location i, an action $a \in \{1,\ldots,m\}$ is selected, a real-valued reward $r(i,a)$ is received, and the next state is determined by the transition probabilities $p(i,a,j)$ for $j \geq 0$. This is repeated indefinitely.

A policy f for controlling this chain is a mapping from the state space $\{0,1,\ldots\}$ to the action space $\{1,\ldots,m\}$, with the meaning that action $f(i)$ is selected whenever the chain is in state i. We shall consider only these so-called stationary deterministic policies. As pointed out below, nothing is gained by considering nonstationary history dependent policies.

Each policy f, along with a rule for starting the process, determines a stochastic process $\{(X_n, a_n) : n \geq 0\}$ where X_n is the state of the chain at time n, and $a_n = f(X_n)$ is the action taken. A policy f^* is called α-discounted optimal if

$$V_{f^*}(i) = \sup_f V_f(i) \quad \text{for all } i,$$

where $0 < \alpha < 1$ is a discount factor and

$$V_f(i) = E_f\Big(\sum_{n=0}^{\infty} \alpha^n r(X_n, a_n) \,|\, X_0 = i\Big).$$

We also define $V_0(i) \equiv 0$,

$$V_n(i) = \sup_f E_f\Big(\sum_{k=0}^{n-1} \alpha^k r(X_k, a_k) \,|\, X_0 = i\Big) \quad \text{for} \quad n \geq 1$$

and

$$V(i) = \sup_f V_f(i).$$

Similarly, a policy f^* is called average optimal if

$$\varphi_{f^*} \quad \sup \varphi_f(i) \quad \text{for all} \quad i$$

where

$$\varphi_f(i) = \liminf_{n \to \infty} n^{-1} E_f\Big(\sum_{k=0}^{n-1} r(X_k, a_k) \,|\, X_0 = i\Big).$$

We shall assume that the rewards are bounded from above, or more generally, that

$$\lim_{n \to \infty} \sup_f E_f\Big(\sum_{k=n}^{\infty} \alpha^k \max\{0, r(X_k, a_k)\} \,|\, X_0 = i\Big) = 0 \quad \text{for each} \quad i.$$

Then the above V's are well-defined and $-\infty \leq V_f(i) \leq V(i) < \infty$ for all i and f, see Schäl (1975). From the theory of dynamic programming--as developed by Bellman, Blackwell, Derman, Howard, Strauch and others, and nicely unified and extended by Hinderer (1970) and Schäl (1975)--we have the following results.

Existence of Stationary Discounted Optimal Policies. Within the class of all history dependent policies, there is a stationary α-discounted optimal policy.

Optimality Criterion. A policy f is α-discounted optimal if and only if

$$U(i,f(i)) = \max_a U(i,a) \quad \text{for all } i,$$

where $U(i,a) = r(i,a) + \alpha \sum_j p(i,a,j)V(j)$.

Optimality Equations. The V_n and V satisfy the equations

$$V_n(i) = \max_a\{r(i,a)+ \alpha \sum_j p(i,a,j)V_{n-1}(j)\} \quad \text{for } n \geq 1, \text{ and}$$

$$V(i) = \max_a\{r(i,a)+ \alpha \sum_j p(i,a,j)V(j)\} \quad \text{for all } i.$$

Value Iteration.

$$V(i) = \lim_{n \to \infty} V_n(i) \quad \text{for all } i.$$

The following results are useful for establishing monotone optimal policies in dynamic programming models, see Serfozo (1977) and Topkis (1978). Here we shall consider the general optimization problem

$$v(i) = \max_a u(i,a) \quad \text{for } i = 0,1,...$$

where $a \in \{1,...,m\}$ and u is a real-valued function. An optimal policy for this problem is defined to be any mapping f from $\{0,1,...\}$ to $\{1,...,m\}$ which satisfies

$$u(i,f(i)) = \max_a u(i,a) \quad \text{for all } i .$$

Note that this is an abstraction of the Optimality Criterion.

Here and throughout this article we use a prime to denote the difference operator with respect to i, namely $w'(i) = w(i+1)-w(i)$. In particular we write $u'(i,a) = u(i+1,a)-u(i,a)$. We also use increasing and decreasing to mean nondecreasing and nonincreasing, respectively.

Proposition 2.1. Let f be the optimal policy defined by

$$f(i) = \max\{a : u(i,a) = \max_{\tilde{a}} u(i,\tilde{a})\} .$$

If $u'(i,a)$ is increasing (decreasing) in a for each i, then f is increasing (decreasing).

Proof. Suppose $u'(i,a)$ is increasing in a, and there is an i such that $f(i+1) < f(i)$. By the definition of $f(i)$ and our suppositions, we have

$$0 \le u(i,f(i)) - u(i,f(i+1)) \le u(i+1,f(i)) - u(i+1,f(i+1)) ,$$

and so $u(i+1,f(i+1)) \le u(i+1,f(i))$. But this contradicts the definition of $f(i+1)$. Thus f must be increasing. A similar argument applies to the decreasing case. ∎

In order to apply Proposition 2.1 when $u(i,a)$ is a function of v, as in a dynamic program, some knowledge of the structure of the value function v may be required. Since v is the upper envelope of $u(\cdot,1),\ldots,u(\cdot,m)$, then v is obviously convex,

increasing or decreasing when all of the $u(\cdot,a)$'s are convex, increasing or decreasing, respectively. The next result describes conditions under which v is concave.

Proposition 2.2. The function v is concave if either of the following conditions hold.

(1) $u'(i,a)$ is increasing in a and $u'(i,1) \geq u'(i+1,m)$ for all i.

(2) $u'(i,a)$ is decreasing in a and $u'(i,m) \geq u'(i+1,1)$ for all i.

Proof. Suppose (1) holds and let f be any optimal policy. Then from (1) we have

$$v'(i) = u(i+1,f(i+1)) - u(i,f(i)) \geq u(i+1,f(i)) - u(i,f(i))$$

$$\geq u'(i,1) \geq u'(i+1,m) \geq u(i+2,f(i+2)) - u(i+1,f(i+2)) \geq v'(i+1) .$$

Thus v is concave. A similar argument shows that v is concave if (2) holds. ∎

3. Monotone Discounted Optimal Policies for Random Walks.

In this section we shall consider a controlled random walk on the nonnegative integers that moves as follows. Upon arriving at a location i, the following events occur:

(a) A pair of probabilities (p_a,q_a) is selected from the set $\{(p_1,q_1),\ldots,(p_m,q_m)\}$. We assume that $0 \leq p_a + q_a \leq 1$ and at least one of these sums is nonzero.

(b) A real-valued reward $r(i,a)$ is received.

(c) The next location of the walk is determined by the transition probabilities

$$p(i,a,i+1) = p_a\,, \quad p(i,a,i-1) = q_a\,, \quad p(i,a,i) = 1-p_a-q_a$$

when $i \geq 1$, and

$$p(0,a,1) = p_a \quad \text{and} \quad p(0,a,0) = 1-p_a\,.$$

This series of events is repeated indefinitely.

The following results describe conditions under which this random walk has increasing or decreasing α-discounted optimal policies. Average optimal policies are discussed in the next section.

Theorem 3.1. Suppose the following conditions hold.

(i) p_a is decreasing and q_a is increasing in a, and $p_1 + q_m \leq 1$.

(ii) $r'(i,a)$ is nonpositive and increasing in a for each i.

(iii) $r'(i,1) \geq r'(i+1,m)$ for each i.

Then there is an increasing α-discounted optimal policy for controlling the random walk.

Theorem 3.2. Suppose the following conditions hold.

(i) p_a is decreasing and q_a is increasing in a.

(ii) $r(i,a)$ is convex and increasing in i for each a.

(iii) $r'(i,a)$ is decreasing in a for each i.

(iv) $\sum_{k=0}^{\infty} \alpha^k \max_a r(k,a) < \infty$.

Then there is a decreasing α-discounted optimal policy for controlling the random walk.

The increasing optimal policy in Theorem 3.1 can be written as

$$f(i) = a \quad \text{if} \quad i_a \le i < i_{a+1},$$

where $0 = i_1 \le i_2 \le \ldots \le i_m \le i_{m+1} = \infty$. This means that if the walk is in location i and $i_a \le i < i_{a+1}$, then (p_a, q_a) is selected to determine the next location (if $i_a = i_{a+1}$, action a is never selected). Since p_a is decreasing and q_a is increasing, the f selects higher q's and lower p's as the location increases. This increases the probability of backward movement of the walk as its location increases, and so the walk tends to stay near zero. Theorem 3.2 describes the opposite situation where it is optimal to decrease the probability of backward movement as the location increases. This tends to push the walk to ∞, accelerating its forward movement as it approaches ∞.

It might appear that condition (i) in both theorems could be replaced by the weaker condition that p_a/q_a is decreasing in a. We feel that this cannot be done, but we do not have a counter-example to justify this. Note that (i) poses no restriction on the (p,q)'s for random walks with a controlled ascent (i.e. $q_1 = \cdots = q_m$) or a controlled descent (i.e. $p_1 = \cdots = p_m$). The action space $\{1,\ldots,m\}$ is assumed to be finite just for simplicity. It could also be a closed interval with the r and p being Borel measurable.

In Theorem 3.1 the assumptions (i)-(iii) insure that the value function V is concave, which is a key ingredient for an increasing

optimal policy. Note that (ii) and (iii) hold if and only if

$$(1)\quad 0 \geq r'(0,m) \geq r'(0,m-1) \geq \cdots \geq r'(0,1) \geq r'(1,m) \geq r'(1,m-1) \geq \cdots \geq r'(1,1) \geq r'(2,m) \geq \cdots$$

This is not too restrictive. It is satisfied when $r(i,a) = g(a)-h(i)$, where h is convex increasing and g has any structure. This is equivalent to $r'(i,a)$ being nonpositive and independent of a for each i. Note that (ii) implies that the rewards are bounded from above; indeed

$$(2)\qquad \sup_{i,a} r(i,a) \leq \max_{a} r(0,a) < \infty .$$

Similarly, in Theorem 3.2 the (i)-(iii) insure that V is convex and that the rewards are bounded from below. Clearly (ii) and (iii) hold if $r(i,a) = g_1(a)+g_2(i)$, where g_2 is convex and increasing. It is also easy to see that (ii) and $\max_a r(i,a) \leq u(i)$, where u is a polynomial, imply (iv).

The proof of Theorem 3.1 is based on the following result. For this we let

$$(3)\qquad f_{n+1}(i) = \max\{a : U_n(i,a) = \max_{\tilde{a}} U_n(i,\tilde{a})\},$$

where the U_n is given by (recall Section 2)

$$(4)\qquad U_n(i,a) = r(i,a) + \alpha \sum_j p(i,a,j)V_n(j)$$

$$= \begin{cases} r(0,a) + \alpha[(1-p_a)V_n(0) + p_a V_n(1)] & \text{for } i = 0 \\ r(i,a) + \alpha[q_a V_n(i-1) + (1-p_a-q_a)V_n(i) + p_a V_n(i+1)] & \text{for } i \geq 1 \end{cases}$$

The f_n can be viewed as the optimal policy for the first step in an n-step random walk.

Proposition 3.3. Under the assumptions of Theorem 3.1, the $V_n(i)$ is concave decreasing and $f_n(i)$ is increasing in i for each $n \geq 1$.

Remark 3.4. Consider an N-step random walk as in Theorem 3.1 with a reward function $r_n(i,a)$ for the n-th step (instead of $\alpha^n r(i,a)$), which satisfies (ii) and (iii) for $1 \leq n \leq N$. Then Proposition 3.3 (with different notation) yields an increasing optimal policy at each of the N steps. A similar statement holds for decreasing policies.

Proof. We shall prove this by induction. The assertion is true for $n = 1$ by Propositions 2.1 and 2.2, since $V_1(i) = \max_a r(i,a)$. Now assume the assertion is true for n. The n+1-th Optimality Equation is

$$V_{n+1}(i) = \max_a U_n(i,a).$$

To prove that V_{n+1} is decreasing, it suffices, since V_{n+1} is the upper envelope of the functions $U_n(\cdot,1),\ldots,U_n(\cdot,m)$, to show

(5) $$U_n'(i,a) \leq 0 \quad \text{for each } a \text{ and } i.$$

And to prove that V_{n+1} is concave and f_{n+1} is increasing, it suffices by Propositions 2.1 and 2.2 to show that

(6) $U_n'(i,a)$ is increasing in a for each i, and

(7) $$U_n'(i,1) \geq U_n'(i+1,m) \quad \text{for each } i.$$

To prove (5)-(7), fix an $i \geq 1$. From (4) it follows that

$$(8) \quad U_n'(i,a) = r'(i,a) + \alpha[q_a V_n'(i-1) + (1-p_a-q_a)V_n'(i) + p_a V_n'(i+1)]$$
$$= r'(i,a) + \alpha[V_n'(i) - q_a V_n''(i-1) + p_a V_n''(i)] .$$

Under the induction hypothesis, the $V_n'(i)$ and $V_n''(i) = V_n'(i+1) - V_n'(i)$ are nonpositive. Then from the first and second lines in (8), and assumptions (i) and (ii), it follows that (5) and (6) hold. The inequality (7) also holds, since (i) and (iii) yield

$$(9) \quad U_n'(i+1,m) - U_n'(i,1) = r'(i+1,m) - r'(i,1) + \alpha[q_1 V_n''(i-1) + (1-p_1-q_m)V_n''(i) + p_m V_n''(i+1)] \leq 0 .$$

Similar arguments prove (5)-(7) for $i = 0$. ∎

We are now ready to establish the existence of an increasing α-discounted optimal policy.

Proof of Theorem 3.1. Consider the policy

$$(10) \quad f(i) = \max\{a : U(i,a) = \max_{\tilde{a}} U(i,\tilde{a})\} ,$$

where $U(i,a)$ is given by (4) with the n's eliminated. By the Optimality Criterion, this f is α-discounted optimal. To complete the proof it suffices, by Proposition 2.1, to show that $U'(i,a)$ is increasing in a for each i. To this end, note that

$$(11) \quad U'(i,a) = \begin{cases} r'(0,a) + \alpha[V'(0) - q_a V'(0) + p_a V''(0)] & \text{for } i = 0 \\ r'(i,a) + \alpha[V'(i) - q_a V''(i-1) + p_a V''(i)] & \text{for } i \geq 1 . \end{cases}$$

By Proposition 3.3 and the Value Iteration $V = \lim_{n\to\infty} V_n$, it follows that V is concave decreasing. Then using (i), (ii), $V'(0) \le 0$ and $V''(i) \le 0$ in (11) we see that $U'(i,a)$ is increasing in a. ▌

Proof of Theorem 3.2. From (ii), (iv) and $P_f(X_k \le i+k \mid X_0 = i) = 1$ we have

$$\lim_{n\to\infty} \sup_f E_f\Big(\sum_{k=n}^{\infty} \alpha^k |r(X_k, a_k)| \mid X_0 = i \Big) \le \lim_{n\to\infty} \sum_{k=n}^{\infty} \alpha^k \max_a |r(i+k, a)| = 0 .$$

Thus the dynamic programming results in Section 2 hold.

Similar to Proposition 3.3 one can show that each n-period value function $V_n(i)$ is convex and increasing in i and $f_n(i)$ is decreasing in i. In the induction argument, the V_n is convex since it is the upper envelope of $U_{n-1}(\cdot,1),\ldots,U_{n-1}(\cdot,m)$ which are clearly convex.

Now consider the α-discounted optimal policy f as defined by (10). Arguing as in Theorem 3.1, using the property that $V(i) = \lim_{n\to\infty} V_n(i)$ is convex increasing, it follows that f is decreasing. ▌

The next result describes some properties of the increasing α-discounted policies in Theorem 3.1. A similar result holds for decreasing policies.

Theorem 3.5. Suppose the random walk satisfies the hypotheses of Theorem 3.1 and let $f(i,\alpha)$ be the increasing α-discounted optimal policy defined by (10).

(a) The $f(i,\alpha)$ is increasing in α for each i.

(b) If there is an α_0 and M such that

$$(12) \qquad r'(i-1,m) \le [\alpha_0(p_{m-1}-p_m+q_m-q_{m-1})]^{-1}\min\{r(i,b)-r(i,a): 1\le a\le b\le m\}$$

for each $i \ge M$, then

$$(13) \qquad f(i,\alpha) = m \quad \text{for each } i \ge M \text{ and } \alpha \ge \alpha_0 .$$

In particular, (13) holds for some α_0 and M if

$$(14) \qquad r(i,a) = g(a)-h(i) \text{ and } h(i) \text{ is strictly convex and increasing.}$$

Remarks. Assertion (a) implies that the set $\{f(\cdot,\alpha) : 0 < \alpha < 1\}$ consists of a finite number of distinct policies. We use this in Theorem 4.2 to obtain an increasing average optimal policy from this set. Assertion (13) implies that the search for an optimal policy from the infinite set of increasing policies can be restricted to a finite set of increasing policies.

Proof. (a) For this part we shall affix an α to the functions defined above to show their dependency on α. Accordingly, we let $f_n(i,\alpha)$ be the optimal solution, as in (3), to the optimization problem

$$V_n(i,\alpha) = \max_a U_{n-1}(i,a,\alpha) \quad \text{for } 0 < \alpha < 1 .$$

We shall first show by induction on n that $f_n(i,\alpha)$ is increasing in α and $V_n'(i,\alpha)$ is decreasing in α for each i and n. This is true for $n=1$, since $V_1(i,\alpha) = \max_a r(i,a)$ and $f_1(i,\alpha)$ are independent of α. Now suppose it is true for n. Then clearly

$$(15)\quad U_n(i,a+1,\alpha)-U_n(i,a,\alpha)=\begin{cases} r(0,a+1)-r(0,a)+\alpha(p_{a+1}-p_a)V_n'(0,\alpha) & \text{if } i=0 \\ r(i,a+1)-r(i,a)+\alpha[(p_{a+1}-p_a)V_n'(i,\alpha) & \\ \qquad +(q_a-q_{a+1})V_n'(i-1,\alpha)] & \text{if } i\geq 1, \end{cases}$$

and the right side is increasing in α for each i and a. Thus by Proposition 2.1, with i replaced by α, it follows that $f_{n+1}(i,\alpha)$ is increasing in α for each i. For a fixed $i\geq 1$ and α, we can write

$$(16)\quad V_{n+1}'(i,\alpha)=U_n(i+1,b,\alpha)-U_n(i,a,\alpha)=r(i+1,b)-r(i,a)+$$
$$+\alpha[p_bV_n'(i+1,\alpha)+(1-p_a-q_b)V_n'(i,\alpha)+q_aV_n'(i-1,\alpha)],$$

where $a=f_{n+1}(i,\alpha)$ and $b=f_{n+1}(i+1,\alpha)$. Since $f_{n+1}(\cdot,\alpha)$ is increasing in α, there is a $\beta>\alpha$ such that $f_{n+1}(i,\beta)=a$ and $f_{n+1}(i+1,\beta)=b$. Using $0\geq V_n'(i,\alpha)\geq V_n'(i,\beta)$, and $1-p_1-q_m\geq 0$ in (16), it follows that $V_{n+1}'(i,\alpha)\geq V_{n+1}'(i,\beta)$. Thus $V_{n+1}'(i,\alpha)$ is decreasing at α for each α and $i\geq 1$. This also follows for $i=0$, by a similar argument. Hence the induction is complete.

Now consider the policy $f(i,\alpha)$ which is optimal for the problem

$$V(i,\alpha)=\max_a U(i,a,\alpha)\qquad 0<\alpha<1.$$

We can write $U(i,a+1,\alpha)-U(i,a,\alpha)$ as in (15) without the n's. By the last paragraph and Value Iteration, we know that $V'(i,\alpha)=\lim_{n\to\infty}V_n'(i,\alpha)$ is decreasing in α, and so $U(i,a+1,\alpha)-U(i,a,\alpha)$ is increasing in α

for each i and a. Thus by Proposition 2.1, it follows that $f(i,\alpha)$ is increasing in α for each i.

(b) Fix an $\alpha \geq \alpha_0$ and $i \geq M$. We also assume $i \geq 1$: the proof is similar for $i = 0$. To prove (13) it suffices to show that

$$U(i,m) \geq U(i,a) \quad \text{for all} \quad a \leq m-1,$$

where U is defined by (4) without the n's. We can write

$$U(i,a) = r(i,a) + \alpha \sum_j p(i,a,j)[r(j,f(j,\alpha))+\alpha W(j)]$$

where

$$W(j) = \sum_k p(j,f(j,\alpha),k)V(k).$$

Letting $a_0 = f(i-1,\alpha)$, $a_1 = f(i,\alpha)$ and $a_2 = f(i+1,\alpha)$, we have

$$(17)\quad U(i,m)-U(i,a) = r(i,m)-r(i,a)+\alpha(p_m-p_a)[r'(i,a_2)+r(i,a_2)-$$

$$-r(i,a_1)+\alpha W'(i)] + \alpha(q_a-q_m)[r'(i-1,a_1)+r(i-1,a_1)-r(i-1,a_0)+\alpha W'(i-1)]$$

We know that $V' \leq 0$, and so

$$W'(i) = p_{a_2}V'(i+1)+(1-p_{a_1}-q_{a_2})V'(i)+q_{a_1}V'(i-1) \leq 0,$$

since $1-p_{a_1}-q_{a_2} \geq 1-p_1-q_m \geq 0$. Similarly $W'(i-1) \leq 0$. Note also that $r(i,a_2) \leq r(i,a_1)$, for if not, then

$$U(i,a_2)-U(i,a_1) = r(i,a_2)-r(i,a_1)+\alpha[(p_{a_2}-p_{a_1})V'(i)+(q_{a_1}-q_{a_2})V'(i-1)] > 0$$

which would contradict the optimality of $a_1 = f(i,\alpha)$. Similarly, $r(i-1,a_1) \leq r(i-1,a_0)$. Using these observations in (17), along with

$$r'(i,a_2) \leq r'(i-1,a_1) \leq r'(i-1,m), \quad p_1 + q_m \leq 1$$

and (12), it follows that

$$U(i,m)-U(i,a) \geq r(i,m)-r(i,a)+\alpha(p_m-p_a+q_a-q_m)r'(i-1,m) \geq 0.$$

This proves (13). Condition (14) clearly implies (12) and hence (13). ▌

4. Monotone Average Reward Policies for Random Walks

If a Markov decision process has discounted optimal policies with a special structure, then it is reasonable that there should be an average optimal policy with the same structure. We shall show that this is true for our random walk. We also present a linear program for computing average optimal policies.

We begin with a result that is useful for obtaining a structured average optimal policy from structured α-discounted optimal policies.

For this we let (X_n, a_n) be a controlled Markov chain as in Section 2. Let Π denote the set of all policies f under which the α-discounted value function $V_f(\cdot)$ is finite for all α in some neighborhood of 1, and the limit

$$\varphi_f(i) = \lim_{n\to\infty} n^{-1} E_f\Big(\sum_{k=0}^{n-1} r(X_k, a_k) \mid X_0 = i\Big)$$

exists for all i, where $-\infty \le \varphi_f(i) < \infty$. The following is an extension of Theorem 1 of Derman (1962).

Proposition 4.1. (a) If f^* is a policy in Π which is discounted optimal for discount factors $\alpha_1, \alpha_2, \ldots$ where $\alpha_n \to 1$, then

$$\varphi_{f^*}(i) = \sup_{f \in \Pi} \varphi_f(i) \quad \text{for all } i. \tag{1}$$

(b) Suppose there is a subset $\Pi^* \subset \Pi$ which contains an α-discounted optimal policy for each α in some interval $[\beta, 1]$, and $\varphi_f(i) \equiv -\infty$ for all $f \notin \Pi$. If Π^* is a finite set, then it contains an average optimal policy.

Proof. Suppose for now that $r \le 0$. We first note that for any $f \in \Pi$,

$$\varphi_f(i) = \lim_{\alpha\to 1}(1-\alpha)V_f(i) \quad \text{for all } i. \tag{2}$$

This follows by the well-known Tauberian theorem [7, p.447] when $\varphi_f(i)$ is finite. And it follows when $\varphi_f(i) = -\infty$, since

$$(1-\alpha)V_f(i) \le (1-\alpha)E_f\Big(\sum_{k=0}^{\nu_\alpha} \alpha^k r(X_k, a_k) \mid X_0 = i\Big) \tag{3}$$

$$\le \nu_\alpha^{-1} E_f\Big(\sum_{k=0}^{\nu_\alpha} \alpha^k r(X_k, a_k) \mid X_0 = i\Big) \to \varphi_f(i) = -\infty \quad \text{as } \alpha \to 1,$$

where ν_α is the integer part of $(1-\alpha)^{-1}$.

Now using (2) we have for any $f \in \Pi$

$$\varphi_f(i) = \lim_{\alpha \to 1}(1-\alpha)V_f(i) \leq \lim_{n \to \infty}(1-\alpha_n)V_{f^*}(i) = \varphi_{f^*}(i) .$$

This proves (1) for nonpositive r. For the general case, let c be an upper bound for r and consider the Markov decision process with rewards $\tilde{r}(i,a) = r(i,a)-c$, transition probabilities $p(i,a,j)$, and average rewards $\tilde{\varphi}_f$. Since $\tilde{r} \leq 0$ and $\tilde{\varphi}_f = \varphi_f - c$, we have

$$\varphi_{f^*}(i) = \tilde{\varphi}_{f^*}(i)+c = \sup_{f \in \Pi} \tilde{\varphi}_f(i)+c = \sup_{f \in \Pi} \varphi_f(i) .$$

This proves part (a). Part (b) follows from (a), since the finite set Π^* must contain a policy which is α_n-discounted optimal for some $\alpha_n \to 1$. ∎

We are now ready to prove the following analogue of Theorem 3.1. A similar result holds for decreasing policies.

Theorem 4.2. Suppose the controlled random walk defined in Section 3 satisfies the following conditions.

(i) p_a is decreasing and q_a is increasing in a, and $p_1+q_m \leq 1$.

(ii) $r'(i,a)$ is nonpositive and increasing in a for each i.

(iii) $r'(i,1) \geq r'(i+1,m)$ for each i.

(iv) $\sum_{k=0}^{\infty} \alpha^k \min_a r(k,a) > -\infty$.

(v) $q_1 > 0$, $p_1 > 0$, $p_m/q_m < 1$ and $r'(i,a) < 0$ for some i and a.

Then there is an increasing average reward policy for controlling the random walk.

Proof. We first establish the existence of average rewards. Let f be any policy, and let

$$\gamma_i = \prod_{k=0}^{i-1} p_{f(k)}/q_{f(k+1)} \quad \text{for } i \geq 1, \tag{4}$$

and

$$N = \inf\{i \geq 0 : p_{f(i)}=0\}.$$

By the standard theory of random walks we have the following cases.

Case 1. The random walk under f is such that $\{0,1,\ldots,N\}$ is a closed class of positive recurrent states and $\{N+1,N+2,\ldots\}$ are transient states if and only if $\sum_i \gamma_i < \infty$. (When $N=\infty$ the latter set is empty.)

Case 2. The random walk under f has all transient or null recurrent states if and only if $\sum_i \gamma_i = \infty$.

In Case 1 we know that

$$\varphi_f(i) = r(0,f(0))\pi_f(0) + \sum_{k=1}^{N} r(k,f(k))\gamma_k \quad \text{for all } i, \tag{5}$$

where

$$\pi_f(i) = \begin{cases} \gamma_i \Pi_f(0) & \text{for } i \geq 1 \\ \left(1 + \sum_{i=1}^{\infty} \gamma_i\right)^{-1} & \text{for } i = 0 \end{cases} \tag{6}$$

is the limiting distribution of the walk. Here $-\infty \le \varphi_f < \infty$, and if $N < \infty$, then $\Pi_f(i) = \gamma_i = 0$ for $i > N$.

In Case 2 we have $\varphi_f(i) = -\infty$ for all i. To see this, let $\nu_i(n)$ denote the number of visits that the random walk makes to state i in n steps. Conditions (ii) and (v) imply that $r(i,a) \downarrow -\infty$ as $i \to \infty$ for each a. Then for any $j > 1$

$$n^{-1} \sum_{k=0}^{n} r(X_k, a_k) = n^{-1} \sum_{i=0}^{j-1} \nu_i(n) r(i, f(i)) + n^{-1} \sum_{i=j}^{\infty} \nu_i(n) r(i, f(i))$$

$$\le B\, n^{-1} \sum_{i=0}^{j-1} \nu_i(n) + r(j, f(j)) n^{-1} \sum_{i=j}^{\infty} \nu_i(n)$$

$$= r(j, f(j)) + n^{-1} \sum_{i=0}^{j-1} \nu_i(n)\,[B-1],$$

where $B = \max_a r(0,a)$. Since each i is transient or null recurrent, then $n^{-1} \nu_i(n) \to 0$ a.s. It follows that

$$\limsup_{n \to \infty} n^{-1} E_f\Big(\sum_{k=0}^{n-1} r(X_k, a_k) \,|\, X_0 = i \Big) \le r(j, f(j)).$$

Letting $j \to \infty$ yields $\varphi_f(i) = -\infty$ for all i.

We have just established the existence of φ_f for each policy f. Also note that (ii) and (iv) imply that

$$V_f(i) \ge \sum_{n=0}^{\infty} \alpha^n \min_a r(i+n, a) > -\infty \quad \text{for any } f,\ i,\ a \text{ and } \alpha.$$

Now let Π^* denote the set of increasing α-discounted optimal policies defined by (10) in Section 3 for $0 < \alpha < 1$. This set is nonempty since (i)-(iii) are the assumptions of Theorem 3.1 which guarantee the existence of such policies. It is also finite by Theorem 3.5. Thus by Proposition 4.1 there is an increasing average optimal policy in Π^*. ∎

In Theorem 4.2, we assumed (v) simply to eliminate some degenerate cases. The $q_1 > 0$ can be relaxed, but more details are involved (cf.(4)). The $p_1 > 0$ along with $q_1 > 0$ just rules out the case in which $p_1 = \ldots = p_m = 0$ and each policy determines a walk that is absorbed at zero. The $p_m/q_m < 1$ eliminates the case in which each f determines a transient or null recurrent walk with $\varphi_f = -\infty$. Here any policy is average optimal. Assumption (iv) insures that V_f is finite-valued, which is needed to apply Proposition 4.1. Clearly (ii) and $\min_a r(i,a) \geq g(i)$, where g is a polynomial, imply (iv).

The random walk we have been considering has an infinite state space, and so we cannot compute optimal policies for it by the standard procedures for finite state processes. Variations of these procedures do apply, however, in some cases as we shall now describe.

Consider a random walk, such as in Theorem 4.2, which has an increasing average optimal policy f. We assume that $q_1 > 0$ and $q_m > p_m$. This insures that each policy determines a positive recurrent random walk. We also assume that f is constant, say equal to m, on the states $\{M, M+1, \ldots\}$ where M is known. This holds, for example, if r satisfies (12) or (14) in Theorem 3.5.

Associated with this walk we let $\pi = \{\pi(i,a) : i \geq 1,\ 1 \leq a \leq m,\ \pi(i,m)=1 \text{ for } i \geq M\}$ denote a randomized policy such that $\pi(i,a)$ is the probability of selecting action a when the walk is in

state i, and m is selected for all $i \geq M$. Under the policy π, the Markov chain (X_n, a_n) is positive recurrent. Letting

$$\nu(i,a) = \lim_{n\to\infty} P_\pi(X_n=i,\ a_n=a \mid X_0=j),$$

the average reward is

$$\varphi_\pi = \sum_{i=0}^{\infty} \sum_{a=1}^{m} \nu(i,a) r(i,a). \tag{7}$$

We can write

$$\nu(i,a) = \nu(i)\pi(i,a) \quad \text{and} \quad \nu(i) = \sum_{a=1}^{m} \nu(i,a), \tag{8}$$

where

$$\nu(i) = \lim_{n\to\infty} P_\pi(X_n=i \mid X_0=j).$$

Since $\pi(i,m) = 1$ for $i \geq M$, then by (6) it follows that

$$\nu(i) = \nu(M)(p_m/q_m)^{i-M} \quad \text{for} \quad i \geq M. \tag{9}$$

Consequently, expression (7) simplifies to

$$\varphi_\pi = \sum_{i=0}^{M-1} \sum_{a=1}^{m} \nu(i,a) r(i,a) + c(M) \sum_{a=1}^{m} \nu(M,a)$$

where

$$c(M) = \sum_{k=0}^{\infty} r(M+k,m)(p_m/q_m)^k. \tag{10}$$

The problem of maximizing the φ_π over π is clearly equivalent to the following linear program:

$$\max_{\nu(i,a)} \sum_{i=0}^{M-1} \sum_{a=1}^{m} \nu(i,a) r(i,a) + c(M) \sum_{a=1}^{m} \nu(M,a)$$

subject to

$$\sum_{a=1}^{m} \nu(0,a) = \sum_{a=1}^{m} [\nu(0,a)(1-q_a) + \nu(1,a)q_a]$$

$$\sum_{a=1}^{m} \nu(j,a) = \sum_{a=1}^{m} [\nu(j-1,a)p_a + \nu(j,a)(1-p_a-q_a) + \nu(j+1,a)q_a]$$

for $1 \le j \le M-1$,

$$\sum_{a=1}^{m} \nu(M,a) = \sum_{a=1}^{m} [\nu(M-1,a)p_a + \nu(M,a)(1-p_a-q_a) + \nu(M,a)(p_m/q_m)q_a]$$

$$\sum_{i=0}^{M-1} \sum_{a=1}^{m} \nu(i,a) + (1-p_m/q_m)^{-1} \sum_{a=1}^{m} \nu(M,a) = 1$$

$$0 \le \nu(0,a) \le \ldots \le \nu(M,a) \le 1 \quad \text{for } 1 \le a \le m.$$

Note that the constraints imply that $\nu(i,a)$ is the limiting distribution of $\{(X_n, a_n)\}$. An optimal solution $\nu(i,a)$ of the linear program, determines the increasing average optimal policy

$$\pi(i,a) = \nu(i,a)\Big(\sum_{a=1}^{m} \nu(i,a)\Big)^{-1} \qquad 0 \le i \le M-1$$

$$\pi(i,m) = 1 \qquad i \ge M.$$

This will be a nonrandom policy when $\nu(i,a)$ is calculated by the simplex algorithm. Note that the exceptionally nice limiting distribution (9) is the key property that allows us to reduce our infinite variable optimization problem to a finite variable problem. This cannot be done for most controlled Markov chains. Even if

the M is not known for sure, the above linear program could be run for a large M and for m equal to 1,2,..., and m. The best of the resulting policies should be close to being average optimal over all policies.

For some random walks with simple reward functions $r(i,a)$, such as polynomials in i, the average reward φ_f in (5) for a monotone policy might be tractable enough to obtain an optimal policy by policy improvement or by other ad hoc arguments. An illustration of this is as follows.

Example 4.3. Consider a random walk with two possible actions that satisfies the following conditions.

(i) $0 < q_1 \le q_2$, $p_1 + q_2 \le 1$, $\rho_1 > \rho_2$ and $\rho_2 < 1$, where $\rho_a = p_a/q_a$.

(ii) $r(i,a) = g(a) - hi$, where (for simplicity) $g(1) > g(2)$ and $h > 0$.

Let

$$D_n = \begin{cases} -n/(1-\rho_1) + (\rho_1-\rho_2)(1-\rho_1^n)(1-\rho_1)^{-2}(1-\rho_2)^{-1} + \rho_2(g(1)-g(2))h^{-1}(\rho_1-\rho_2)^{-1} \\ \qquad\qquad \text{if } \rho_1 \neq 1 \\ -n^2 - n(1+\rho_2)(1-\rho_2)^{-1} + 2\rho_2(g(1)-g(2))h^{-1}(1-\rho_2)^{-1} \quad \text{if } \rho_1 = 1, \end{cases}$$

and

$$n^* = \min\{n : D_n \le 0\}.$$

Then an average optimal policy is to select (p_1,q_1) when the walk is below state n^* and select (p_2,q_2) otherwise.

To verify this assertion, let φ_n and $\pi_n(i)$ denote the average reward and limiting distribution for the walk under the policy of using (p_1,q_1) when and only when the walk is below state n. Then from (4)-(6) and some lengthy algebraic manipulations one can show that

$$\varphi_{n+1} - \varphi_n = \begin{cases} \rho_1^{n-1} \pi_n(0)\pi_{n+1}(0)(\rho_1-\rho_2)(1-\rho_2)^{-1}hD_n & \text{if } \rho_1 \neq 1 \\ 1/2\, \pi_n(0)\pi_{n+1}(0)hD_n & \text{if } \rho_1 = 1, \end{cases}$$

and that D_n is strictly decreasing. This factorization implies that ϕ_n has a global maximum at n^*. This and Theorem 4.2 yield the assertion.

5. Random Walks with State Dependent Transitions

We have been discussing a random walk whose step size is determined by a pair (p_a,q_a) of probabilities selected from a set which is independent of the location of the walk. Results similar to those in Sections 3 and 4 also hold for random walks whose set of possible actions may depend on the location of the walk. We illustrate this here by presenting analogues of Theorems 3.1 and 3.2.

Consider a random walk on $\{0,1,\ldots\}$ that moves as follows. Upon arriving at a location i the following events occur:

(a) A pair of probabilities $(p(i,a),q(i,a))$ is selected from the set $\{(p(i,1),q(i,1)),\ldots,(p(i,m),q(i,m))\}$.

(b) A reward $r(i,a)$ is received.

(c) The next location of the walk is determined by the transition probabilities

$$p(i,a,i+1) = p(i,a), \quad p(i,a,i-1) = q(i,a), \quad p(i,a,i) = 1-p(i,a)-q(i,a),$$

when $i \geq 1$, and

$$p(0,a,1) = p(0,a) \quad \text{and} \quad p(0,a,0) = 1-p(0,a) .$$

This series of events is repeated indefinitely.

For the following result, we let

$$d(0,a) = -p(0,a) \quad \text{and} \quad d(i,a) = q(i,a)-p(i,a) \quad \text{for} \quad i \geq 1 .$$

Theorem 5.1. Suppose $p(i,1) \geq \ldots \geq p(i,m)$, $q(i,1) \leq \ldots \leq q(i,m)$, $d'(i,1) \leq \ldots \leq d'(i,m)$, and $p(i,a) + q(i+1,a) \leq 1$ for each i and a.

(a) If $r'(i,1) \leq \ldots \leq r'(i,m) \leq 0$, $r'(i,1) \geq r'(i+1,m)$, $d'(i,1) \geq d'(i+1,m)$, and $p(i+1,1) + q(i+2,m) - p'(i+1,m) \leq 1$ for each i, then there is an increasing α-discounted optimal policy for controlling the random walk.

(b) If $d(i,a)$ is concave in i and $r(i,a)$ is convex and increasing in i for each a; and for each i and a, $r'(i,1) \geq \ldots \geq r'(i,m)$,

$$\sum_{k=0}^{\infty} \alpha^k \max_a r(k,a) < \infty , \text{ and } p(i+1,a) + q(i+2,a) - p'(i+1,a) \leq 1 ,$$

then there is a decreasing α-discounted optimal policy for controlling the random walk.

Proof. The assertions follow by arguments as used in proving Theorems 3.1 and 3.2. The key idea is to use the following factorizations in the induction arguments (these are for $i \geq 1$):

$$U_n'(i,a) = r'(i,a)+\alpha\{[1-p(i,a)-q(i+1,a)]V_n'(i)+p(i+1,a)V_n'(i+1)+q(i,a)V_n'(i-1$$

$$U_n'(i+1,m)-U_n'(i,1) = r'(i+1,m)-r'(i,1)+\alpha\{p(i+2,m)V_n''(i+1)+q(i,1)V_n''(i-1)$$

$$+ [1-p(i+1,1)-q(i+2,m)+p'(i+1,m)]V_n''(i)+[d'(i,1)-d'(i+1,m)]V_n'(i)\},$$

and

$$U_n''(i,a) = r''(i,a)+\alpha\{p(i+2,a)V_n''(i+1)+q(i,a)V_n''(i-1)-d''(i,a)V_n'(i)$$

$$+ [1-p(i+1,a)-q(i+2,a)+p'(i+1,a)]V_n''(i)\}.$$

6. Controlled Birth and Death Processes

In this section we shall consider a controlled birth and death process on the nonnegative integers that moves as follows. Upon arriving at a state i, the following events occur.

(a) A pair of birth-death parameters (λ_a,μ_a) is selected from the set $\{(\lambda_1,\mu_1),\ldots,(\lambda_m,\mu_m)\}$ where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m > 0$ and $0 < \mu_1 \leq \mu_2 \leq \ldots \leq \mu_m$. We also assume $\lambda_m < \mu_m$ when considering average optimal policies.

(b) The process remains in state i for a random time which has an exponential distribution with parameter

$$\lambda(i,a) = \begin{cases} \lambda_a & \text{if } i = 0 \\ \lambda_a + \mu_a & \text{if } i \geq 1. \end{cases}$$

Then the process jumps to a neighboring state according to the transition probabilities $q(0,a,1) = 1$ for $i = 0$, and

$$q(i,a,i+1) = \lambda_a/(\lambda_a+\mu_a), \text{ and } q(i,a,i-1) = \mu_a/(\lambda_a+\mu_a) \text{ for } i \geq 1.$$

(c) A cost is incurred at a rate $c(a)+h(i)$ per unit time during the sojourn in i, where h is convex and increasing. A nonnegative reward R is also received for a birth (when the process jumps from i to $i+1$). This series of events is repeated indefinitely.

We first show that this continuous time controlled birth and death process is equivalent to a discrete time controlled random walk. Then we use Theorems 3.1 and 4.2 to establish the existence of increasing discounted and average optimal policies for the birth and death process. We also discuss the M|M|1 queue with controlled arrival and service rates.

We begin by introducing some more notation. A policy f for successively choosing the birth-death parameters is defined to be a mapping from the state space $\{0,1,\ldots\}$ to the action space $\{1,\ldots,m\}$. Each policy f, along with a rule for starting the process, determines a continuous time birth and death process whose birth-death parameters in state i are $(\lambda_{f(i)},\mu_{f(i)})$. We let Y_n and T_n denote the n-th state of the process, and the time at which the process jumps to state Y_n, respectively. The discounted reward for the process if given by

$$W_f(i) = E_f\Big(\sum_{n=0}^{\infty} e^{-\beta T_n} g_\beta(Y_n,a_n) \,|\, X_0=i\Big), \quad \text{where } a_n = f(Y_n),$$

$\beta > 0$ is a continuous time discount factor, and $g_\beta(i,a)$, the discounted gain in a sojourn, is

$$g_\beta(i,a) = E_f\Big(e^{-\beta T_1} R\delta(Y_1,i+1) - \int_0^{T_1} (c(a)+h(i))e^{-\beta t}\,dt \,|\, Y_0=i,\ a_0=a\Big)$$

$$= (\lambda_a R - c(a) - h(i))/(\beta+\lambda(i,a)).$$

Here T_1 is the exponential sojourn time in state i, the $\delta(i,j) = 1$ or 0 according as $i=j$ or $i\neq j$, and a reward ρ received at time t has a value $\rho e^{-\beta t}$. Similarly, the average reward for the process is

$$\psi_f(i) = \liminf_{t\to\infty} t^{-1} E_f\Big(\sum_{n=0}^{N_t} g_0(X_n, a_n) \mid X_0=i\Big),$$

where $N_t = \sup\{n : T_n \le t\}$ is the number of jumps in time t. A policy f^* is called β-discounted optimal if

$$W_{f^*}(i) = \sup_f W_f(i) \quad \text{for all } i,$$

and f^* is called average optimal if

$$\psi_{f^*}(i) = \sup_f \psi_f(i) \quad \text{for all } i.$$

The following result is a special case of the equivalence between continuous and discrete time Markov decision processes in Lippman (1975) and Serfozo (1979).

Theorem 6.1. Consider the birth and death process defined above, and consider the random walk defined in Section 3 with

$$p_a = \lambda_a/\Lambda, \text{ and } q_a = \mu_a/\Lambda \text{ where } \Lambda = \lambda_1 + \mu_m,$$

$$r(i,a) = (\lambda_a R - c(a) - h(i))/(\beta+\Lambda),$$

and $\alpha = \Lambda/(\beta+\Lambda)$. A policy is β-discounted (average) optimal for the birth and death process if and only if it is α-discounted (average) optimal for the random walk.

Proof. Note that the transition probabilities and rewards of the two processes are such that

$$p(i,a,j) = \begin{cases} \lambda(i,a)q(i,a,j)/\Lambda & \text{if } i \neq j \\ 1 - \lambda(i,a)/\Lambda & \text{if } i = j, \end{cases}$$

and

$$r(i,a) = g_\beta(i,a)(\beta+\lambda(i,a))/(\beta+\Lambda) .$$

Then from [16], it follows for any policy f that

$$W_f(i) = V_f(i) \quad \text{and} \quad \psi_f(i) = \Lambda\varphi_f(i) \quad \text{for all } i .$$

This yields the assertion. ∎

We are now ready for our main result.

Theorem 6.2. There exist increasing β-discounted and average optimal policies for controlling the birth and death process.

Proof. The random walk described in Theorem 6.1 clearly satisfies the assumptions of Theorem 3.1. Here $r'(i,a) = -h'(i)(\beta+\Lambda)^{-1}$. Then there exists an increasing α-discounted optimal policy for the random walk, and by Theorem 6.1 this policy is also β-optimal for the birth and death process. Similarly, the existence of an increasing average optimal policy follows by Theorems 4.2 and 6.1. ∎

We now return to the queueing example that we briefly discussed in Section 1.

Example. Consider an M|M|1 queue whose arrival and service rates are controlled as follows. At each service completion or customer

arrival, the number of customers in the system is observed. Based on this number, an arrival and service rate pair (λ_a,μ_a) is selected from the set $\{(\lambda_1,\mu_1),\ldots,(\lambda_m,\mu_m)\}$. A cost $c(a)$ per unit time is charged for using (λ_a,μ_a), a cost $h(i)$ is charged for holding i customers in the system for one time unit, and a reward R is received from each entering customer. Then under the assumptions in Theorem 6.2 on the (λ_a,μ_a)'s and costs, there exist increasing β-discounted and average optimal policies for controlling the queue.

It is interesting to note that increasing policies may not be optimal for this queueing process if its state space is finite. We have found such examples for both discounted and average rewards, also see [3]. The reason is that the rejection of customers arriving to a full queue limits the holding cost, and so a service and arrival rate that is optimal for a given state may not be economical in a higher state to reduce the queue. Also note that we have tacitly assumed that there is no cost for changing the arrival and service rates. When such costs are involved, the analysis is more complex. This problem has only been solved for the case when the arrival rate is fixed and the service rate can be set at 0 or at a fixed rate $\mu > 0$, see Deb (1976).

We have been considering a birth and death process with a special reward structure just for the sake of simplicity. Theorem 6.2, and an analogue for decreasing policies, also hold for processes with more general rewards (as in Theorems 3.1 and 3.2), and with state-dependent action spaces (as in Theorem 5.1, an example is in [1]). Because of the equivalence in Theorem 6.1, the computation

of optimal policies for the birth and death process is the same as that for random walks.

Acknowledgement. I thank the referee for pointing out two errors in my original Theorems 3.5 and 4.1, and for his suggestions on improving my first draft.

References

[1] Albright, S.C. and W. Winston (1977) A birth death model of advertising. Technical report. School of Business, Indiana University.

[2] Beja, A. and A. Teller (1975) Relevant policies for Markovian queueing systems with many types of service. Management Science, 21, 1049-1054.

[3] Crabill, T. (1972) Optimal control of a service facility with variable exponential service time and constant arrival rate. Management Science, 18, 560-566.

[4] Crabill, T., Gross, D. and M. Magazine (1977) A classified bibliography of research on optimal design and control of queues. Operations Research, 25, 219-232.

[5] Deb, R. (1976) Optimal control of batch service queues with switching costs. Adv. Appl. Prob., 8, 177-194.

[6] Derman, C. (1962) On sequential decisions and Markov chains. Management Science, 9, 16-24.

[7] Feller, W. (1971) Introduction to Probability and Its Applications Vol. II, 2nd Ed., John Wiley, New York.

[8] Heyman, D. (1968) Optimal operating policies for M|G|1 queueing systems. Operations Research, 16, 362-382.

[9] Hinderer, K. (1970) Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter. Lecture Notes in Operations Research and Mathematical Sciences #33, Springer-Verlag, New York.

[10] Lippman, S. (1975) Applying a new device in the optimization of exponential queueing systems. Operations Research, 23, 687-710.

[11] Low, D. (1974a) Optimal dynamic pricing policies for an M|M|s queue. Operations Research, 22, 545-561.

[12] Low, D. (1974b) Optimal pricing for an unbounded queue. IBM Journal of Research and Development, 18, 290-302.

[13] Sabeti, H. (1973) Optimal selection of service rates in queueing with different costs. J. Operations Research Soc. of Japan, 16, 15-35.

[14] Schäl, M. (1975) Conditions for optimality in dynamic programming for the limit of n-stage optimal policies to be optimal. Z. Wahrscheinlichkeitstheorie verw. Geb. 32, 179-196.

[15] Serfozo, R. (1977) Monotone optimal policies for Markov decision processes. Stochastic Systems: Modeling Identification and Optimization, II. Mathematical Programming Study 6, 202-216, North-Holland, New York.

[16] Serfozo, R. (1979) An equivalence between continuous and discrete time Markov decision processes. Operations Research (to appear).

[17] Sobel, M.J. (1969) Optimal average cost policies for a queue with start-up and shut-down costs. Operations Research, 17, 145-162.

[18] Sobel, M.J. (1974) Optimal operation of queues. In Mathematical Methods in Queueing Theory. Lecture notes in Economics and Mathematical Systems, Vol. 98 (Springer, New York) pp.231-261.

[19] Stidham, S. and N.U. Prabhu (1974) Optimal control of queueing systems. In Mathematical Methods in Queueing, Lecture notes in Economics and Mathematical Systems, Vol. 98. (Springer, New York), pp.263-294.

[20] Topkis, D. (1978) Minimizing a submodular function on a lattice. Operations Research, 26, 305-321.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Date Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFOSR-TR-78-1345	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): OPTIMAL CONTROL OF RANDOM WALKS, BIRTH AND DEATH PROCESSES AND QUEUES	5. TYPE OF REPORT & PERIOD COVERED: Interim rept.
6. PERFORMING ORG. REPORT NUMBER	
7. AUTHOR(s): Richard F. Serfozo	8. CONTRACT OR GRANT NUMBER(s): AFOSR-74-2627, NSF-ENG77-13653
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Syracuse University, Department of Industrial Eng. & Operations Res., Syracuse, New York 13210	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS: Air Force Office of Scientific Research/NM, Bolling AFB, Washington, DC 20332	12. REPORT DATE: 1978
	13. NUMBER OF PAGES: 38
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): 40p.	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Markov decision processes
Controlled random walks
Birth and Death processes
Control of Queues

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

We first consider a random walk on the nonnegative integers whose steps are controlled as follows. Upon arriving at a location i, a pair or probabilities (p,q) is selected from a given set, a reward r(i,p,q) is received, and the next step takes the walk to locations i+1, i-1, or i, with respective probabilities p, q and 1-p-q (q=0 when i=0). This is repeated indefinitely. A rule for successively selecting the probabilities (p,q) is a control policy. We identify conditions on the rewards and probabilities under which

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

409 184

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Date Entered)

20. Abstract continued

there exist monotone optimal policies, for discounted and average rewards, and we show how to compute some of these policies. For example, in certain settings it is optimal to increase the tendency of backward movement of the walk as its location increases. We also present similar results for controlling the parameters of a birth and death process and an M/M/1 queue.